Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3226–3231
Marseille, 11–16 May 2020

Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words

Alina Maria Ciobanu, Liviu P. Dinu, Laurentiu Zoicas
University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro, l_zoicas@yahoo.fr

Abstract

Producing related words is a key concern in historical linguistics. Given an input word, the task is to automatically produce either its proto-word, a cognate pair or a modern word derived from it. In this paper, we apply a method for producing related words based on sequence labeling, aiming to fill in the gaps in incomplete cognate sets in Romance languages with Latin etymology (producing Romanian cognates that are missing) and to reconstruct uncertified Latin words. We further investigate an ensemble-based aggregation for combining and re-ranking the word productions of multiple languages.

Keywords: cognates, word production, sequence labeling, multilinguality

1. Introduction

The transition from Latin to Romance languages was followed by a series of vocabulary transformations, whose results are found in all Romance languages (Sala, 1998). The main trend was the simplification of the vocabulary, which consisted mainly in eliminating archaisms and anomalies in favor of regular norms (Sala, 1998), reducing synonyms and deleting sense nuances. These tendencies have led to a diminishing of the vocabulary, making every modern Romance language borrow from Latin about the same number of words – around 2,000. Out of these 2,000 words, about 500 were transmitted to all Romance languages (generally common, important, high-frequency terms), but most of them were preserved only in a subset of the Romance languages. There are, for example, words preserved only in the lateral areas of the Roman Empire, Romanian and the Ibero-Romance languages, such as *frumos* (Ro) - *hermoso* (Es) (meaning *beautiful*). Other words were preserved in only one language (for example, there are about 100 words preserved only in Romanian).

On the other hand, there are about 200 words that evolved from Latin in all Romance languages, but are not found in Romanian. The corresponding Romanian words have probably entered the language at a later stage, as borrowings from other languages. Out of these, many are common words and it is unlikely that they were not part of the Romanian lexicon. Fischer (1985) identified a number of causes that led to the disappearance of these words and to their substitution with words with different etymologies: external causes, of socio-economic nature, the change of people's occupations, the interruption of Romania's contact with the Western world, the development of the Romanian language away from the Romance kernel, and so on. For some fundamental words there are no unanimously accepted explanations. For example, there are no Latin forms for *a iubi* (*to love*), *drag* (*dear*), *cocoş* (*rooster*).

In this paper we propose a computational approach (the first of this type, to the best of our knowledge) to reconstruct these words, starting from cognate sets (cognates in multiple languages) with Latin etymology that are present in all the Romance languages but whose form in Romanian does not exist (according to Reinheimer Ripeanu (2001)). As a secondary research problem, we use the same methodology to automatically reconstruct unattested Latin words (artificially reconstructed by linguists and domain experts), starting from cognate sets in Romance languages.

One of the benefits of this computational approach is that a lot of manual work is spared (provided, of course, that the system takes into account all the transformations through which popular Latin passed on its way to Romanian – or to other languages). In the recent years, a series of articles proposed computational approaches to identifying related words and reconstructing proto-words (Kondrak, 2000; List et al., 2017; Bouchard-Côté et al., 2009; Rama et al., 2018; List, 2019; Ciobanu and Dinu, 2018; Ciobanu and Dinu, 2019), as an alternative to classical comparative reconstruction (Fox, 1995; Campbell, 1998; Weiss, 2015).

Another advantage of the proposed method is that it is possible to reach forms that, non-existent in Romanian, could exist or have existed, as ancient borrowings, in neighboring linguistic spaces, such as Bulgarian, Serbian or Hungarian. It is well-known the case of words – e.g., *borcan* (*jar*) – that the Bulgarians report as borrowed from Romanian, while the Romanians consider them borrowings from Bulgarian; of course they could be substratum elements – Thraco-Dacian, Balkan – but it is not excluded that they belong to the Latin layer.

Last but not least, the approach can serve as a model for processing, in a similar manner, some elements of superstratum (Slavic, Hungarian, German, possibly Turkish or Greek), to understand if these are indeed borrowings or, on the contrary, indigenous words borrowed by the geographically neighboring languages (perhaps very old words, from the religious vocabulary, such as *a mântui* (*to redeem*), of Hungarian origin, according to our dictionaries).

The paper is organized as follows: after an introduction in Section 1, in Section 2 we adapt the method of reconstructing missing cognates starting from a method for identifying proto-words (Ciobanu and Dinu, 2018). In Section 3 we discuss missing Romanian cognates based on the dataset proposed by Reinheimer Ripeanu (2001), we reconstruct them automatically and analyze the results. In Sec-

tion 4 we discuss automatically reconstructing unattested Latin words, which have been previously artificially reconstructed, and finally, in Section 5, we draw conclusions and discuss future work.

2. Producing Related Words

We address the task of automatically producing related words using a method based on sequence alignment and sequence labeling (Ciobanu and Dinu, 2018). We aggregate results from multiple source languages using an ensemble-based aggregation method.

2.1. Sequence Alignment

To align pairs of words, we use the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970),[1] with words as input sequences and a basic substitution matrix, which gives equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *è* are matched).

2.2. Sequence Labeling

To learn changes in spelling and to predict the form of the words in the target language, we use a sequeunce labeling method, namely conditional random fields (CRFs) (Lafferty et al., 2001), using words in the source language as input sequences, their characters as tokens and character n-grams as features. For each character in the target word (after the alignment), the associated label for the CRF system is the character which occurs on the same position in the source word. In the case of insertions, we add the new character to the previous label, because there is no input character in the source language to which we could associate the inserted character as label. We account for affixes separately: for each input word, we add two more characters B and E, marking the beginning and the end of the word. The characters that are inserted in the target word at the beginning or at the end of the word are associated to these special characters. In order to reduce the number of labels, we replace the label with $\star$ for input tokens that are identical to their labels. We used the sequence labeling implementation provided by the Mallet toolkit (McCallum, 2002).

To improve the performance and to take advantage of the information provided by multiple languages, we apply and evaluate an ensemble-based aggregation method that has proven successful in the past (Ciobanu and Dinu, 2018).

Using this methodology, we propose two experiments that illustrate the applicability of word production in historical linguistics and provide insights into the evolution of the Romance languages.

3. Producing Missing Cognates

Having incomplete cognate sets in Romance languages, where the Romanian word is missing, our goal is to automatically produce the Romanian word from the other languages.

[1]The algorithm proposed by (Needleman and Wunsch, 1970) outperformed, in preliminary experiments, the algorithm proposed by Bhargava and Kondrak (2009)

For example, the Latin word *bellus* (meaning *beautiful*) evolved in French (*beau*), Spanish (*bello*), Portuguese (*belo*) and Italian (*bello*), but not in Romanian (according to Reinheimer Ripeanu (2001)).

3.1. Data and Experimental Setup

We ran experiments on cognate sets in Romance languages (Romanian, Italian, French, Spanish, Portuguese) with Latin common ancestors.

We trained the sequence labeling system on cognate sets from the dataset proposed by Ciobanu and Dinu (2014), having Romanian as the target language. We used 2,315 cognate sets for training and 772 for development. We tested the model on the dataset proposed by Reinheimer Ripeanu (2001), which contains 1,102 cognate sets. Out of these, only 372 cognate sets are complete (that is, report a cognate for each Romance language). For all the others, at least one cognate is missing, as follows: Romanian cognates are missing in 493 cognate sets, Italian cognates are missing in 188 cognate sets, French cognates are missing in 245 cognate sets, Spanish cognates are missing in 238 cognate sets and Portuguese cognates are missing in 212 cognate sets.

We focus on Romanian cognates, since they are missing in most cases. Out of the 493 cognate sets in which Romanian cognates are missing, in 235 cognate sets they are the only missing cognates, while in the others, cognates are missing in more languages. We run experiments of reconstructing missing Romanian cognates on the 235 cognate sets where cognates are provided in the Western Romance languages (Portuguese, Spanish, Italian, French), together with their Latin common ancestors, but where Romanian cognates are missing. As future work, we intend to experiment, in turn, with each of the other languages as target language (that is, reconstructing missing cognates in Italian, French, Spanish and Portuguese).

In Table 1 we provide sample cognate sets from both datasets, training and testing. The sample from the former dataset are complete, while the samples from the latter dataset are missing the Romanian cognate, which our system will aim to reconstruct.

We use 3-grams as features and 50 training iterations for the sequence labeling system. On the development dataset, the systems trained on each language obtained the following top-10 accuracy: Spanish 61%, French 62%, Italian 62%, Portuguese 57%, Latin 65%.

The system produces n-best lists of productions. For this task, we perform a very simple rule-based post-processing to correct the results, using the following two rules:

1. We replace the *iă* diphthong with *ie* (since the former does not exist in Romanian).
2. We replace double consonant with single consonants (for example, *ll* becomes *l*).

3.2. Results and Discussion

For the words in the test set (235 cognate sets missing the Romanian cognate), where an automatic evaluation is not possible because a gold standard does not exist, we evaluate and analyze the results through linguistic insights.

Latin	Romanian	Italian	French	Spanish	Portuguese
attractio	atracţie	attrazione	attraction	atracción	atração
láncea	lance	lancia	lance	lanza	lança
orthographia	ortografie	ortografia	orthographie	ortografía	ortografia
physica	fizică	fisica	physique	física	física
vehiculum	vehicul	veicolo	véhicule	vehículo	veículo
bellus	– ? –	bello	beau	bello	belo
cinctura	– ? –	cintura	ceinture	cintura	cintura
extraneus	– ? –	strano	étrange	extraño	estranho
lectus	– ? –	letto	lit	lecho	leito
mercatus	– ? –	mercato	marché	mercado	mercado

Table 1: Sample cognate sets from training and testing datasets.

Starting with the Latin proto-words of these cognate sets and using the method from Section 2 that applies possible transformations undergone by the lexical elements during the evolution towards modern Romance languages, the system generated several word forms that include (or may include) words belonging to the Romanian vocabulary.
This operation was repeated using the same 235 cognate sets, but using the Italian words as the source, with the purpose of producing real or virtual Romanian cognates.[2] We chose Latin as the origin language and Italian as the closest phonetic language to Romanian.
We grouped the results for Latin→Romanian and Italian→Romanian in five types:

1. **Real cognates** – words that exist in Romanian, either as an internal effect – for example, regressive derivatives from verbs: *sărut* (*kiss*), *cânt* (*song*), either as an effect of the process of relatinization of Romanian (initiated in the first quarter of the 19th century) through massively borrowing words from the Romance languages – such as *amic* (*friend*), *insulă* (*island*), *a naviga* (*to navigate*). In Romanian, borrowing from French and Italian continued throughout the 20th century – with words such as *consuetudine* (*custom*), *veritate* (*truth*).

2. **Nuanced real cognates** – words that exist in Romanian, but whose identification was made only after introducing new criteria that were initially disregarded (this refers to some regular phonetic transformations, the rhotacization of intervocalic *-l-*, the regular change of the pitch of some vowels). For example, *concepe*, produced by the system for the Latin word *concipere* (*to conceive*).

3. **Virtual cognates** – words that could exist or could have existed in Romanian if speakers would feel the need for them. It is possible that some of these words actually existed in ancient Romanian or in the old Daco-Romanian. However, as there is no record in this regard, we can only make conjectures.

4. **Nuanced virtual cognates** – words that, similarly to those from the second category, required the introduction of the same new criteria that were initially disregarded. For example, *morină*, produced by the system for the Latin word *morinum* (*mill*).

5. **Inexistent** – words that do not fit in any of the previous categories.

For this classification, we took into account the first 10 productions of the n-best lists produced by the system.
In Table 3 we provide examples of automatically reconstructed cognates from all of the above types.
We make the following remarks about the virtual cognates, the words that are not present in Romanian:

- The analysis started from two datasets compiled by Reinheimer Ripeanu (2001) and Fischer (1985) in which nouns are in the nominative case and not the accusative or ablative ones (much more used in speaking); as a consequence, the system did not generate *voce* from the Latin word *vox* (*voice*) but from the Italian one *voce*. It is a deficiency we should eliminate in the next stage.

- The absence of many Romanian possible cognates has historical causes: their semantic content is found in Romanian in other words either of Latin origin – *piele* (*skin*) from *pellis*, and not from *corium*, or of non-Latin origin – *vrăjmaş*/*duşman* (*enemy*) for *inimic/inamic*. Yet, the process of relatinization aimed precisely the "repopulation" of Romanian with words of Latin origin; this is how *amic* (*friend*) occurred besides *prieten* and *amor* (*love*) besides *iubire* or *dragoste*.

We report the results of the manual evaluation (linguistic insights) in Table 2. When producing Romanian words from Latin, a real or virtual cognate was obtained for about 80%

[2]The produced Romanian cognates are available at `http://nlp.unibuc.ro/resources`. We provide 10-best lists of productions from Italian and Latin.

Type of cognates	Latin	Italian
Real	82 (34.8%)	72 (30.6%)
Nuanced real	12 (5.1%)	11 (4.6%)
Virtual	69 (29.3%)	32 (13.6%)
Nuanced virtual	28 (11.9%)	11 (5.1%)
Inexistent	51 (21.7%)	111 (47.2%)

Table 2: Manual evaluation results for producing missing cognates in Romanian from Italian and Latin.

Source word	Productions	Type of cognates
camera	**cameră**, camer, camera, caeră, camere	Real
durus	**dur**, dură, duru, du, durus	
pratum	**prat**, pra, pratu, prată, prt	
ratio	**raţie**, raţiune, ratie, aţie, ratiune	
tingere	**tinge**, tingere, tinger, tine, tin	
amare	**ama**, am, amă, amar, amare	Virtual
bellus	**bel**, belă, belu, be, el	
duplus	dupl, dup, duplă, **duplu**, dpl	
murus	**mur**, muru, mură, mu, ur	
tardare	**tarda**, tardar, tărda, tara, trda	
cantus	cant **(cânt)**, can, cantă, cănt, cantute	Nuanced real
concipere	concipe **(concepe)**, concipere, conciper, concipa, concip	
iustus	iust **(just)**, iut, iustă, iustute, iustu	
tendere	tende **(tinde)**, tendere, tender, tenda, ten	
velum	vel **(văl)**, velu, ve, el, velă	
attendere	atende **(atinde)**, atendere, atender, atenda, aten	Nuanced virtual
calor	cal, calor, calr, calo, calore **(caloare)**	
gelare	gela **(gera)**, gel, gelă, gelar, gelare	
iungere	iunge **(junge)**, iungere, iunger, iunga, iung	
lectus	lect **(lept)**, lectă, lec, ect, lectu	
alacer	alacer, alac, alacere, alacert,	Inexistent
festadies	festad, festadie, festadies, festadiete, festadi	
medietas	medietate, edietate, medetate, medietat, mdietate	
pater	pater, pat, patr, pator, patee	
profectus	profect, profectă, profec, proect, profectu	

Table 3: Examples of automatically reconstructed Romanian cognates from Latin words. The correct productions are highlighted in bold.

of the test words, while from Italian the percentage was lower (about 53%). Out of the real cognates, 62 were obtained on the first position in the n-best list for Italian and 72 for Latin. This means that, for Latin, summing up the values for the first four categories, we were able to reconstruct the missing Romanian words in 78.3% of the cases. In terms of accuracy, the correct word was on the first position in the productions list in 65% of the cases, and in 75% of the cases it was among the first 5 productions (top-5 accuracy). For Italian, the performance was lower: the inexistent words cover 47% of the cases.

We then aggregated the results obtained from the two languages (Italian and Latin), using ensembles, as described in Section 2, and the results were comparable to those

Latin	Romanian	Italian	French	Spanish	Portughese
aramen	aramă	rame	airain	alambre	arame
ceresia	cireaşă	ciliegia	cerise	cereza	cereja
consutura	cusătură	costura	couture	costura	costura
excappare	scăpa	scappare	échapper	escapar	escapar
expaventare	spăimânta	spaventare	épouvanter	espantar	espantar

Table 4: Examples of cognate sets with unattested Latin proto-words.

Unattested Latin word	Productions
accaptare	accattare, accactare, **accaptare**, acattare, accattari
affumare	**affumare**, aphfumare, abfumare, aefumare, afphumare
novius	novium, **novius**, noivus, novioe, novio
putrire	**putrire**, putrere, putriris, pucrire, putrir
rendere	**rendere**, rindere, rendire, render, rentire

Table 5: Examples of automatically reconstructed unattested Latin proto-words. The correct productions are highlighted in bold.

obtained from Latin (76% of the words could be reconstructed): 84 real cognates, 16 nuanced real cognates, 56 virtual cognates, 22 nuanced virtual cognates and 57 inexistent.

We believe that the results are promising, taking into account the fact that the method is fast and provides a filtering tool for domain experts.

4. Reconstructing Unattested Latin Words

Having modern words in multiple sister languages, where their common Latin etymon is unattested, our goal is to automatically produce the Latin word from which they evolved.

4.1. Data and Experimental Setup

We trained the sequence labeling system on cognate sets from the dataset proposed by Ciobanu and Dinu (2014), having Latin as the target language. We used 1,930 cognate sets for training, and 644 for development. We ran a grid search on the development dataset to determine the optimal size of the window for extracting features and the optimal number of training iterations. We applied the trained model on a list of 63 cognate sets that are present in Romance languages (Portuguese, Spanish, Italian, French, Romanian), but whose common Latin ancestor is unattested (that is, the Latin word is provided, but marked as unattested). The test dataset is extracted from the dataset proposed by Reinheimer Ripeanu (2001). Out of the 63 cognate sets with unattested Latin ancestors, only a few are complete; for most of them, at least one cognate is missing.

In Table 4 we provide examples of complete cognate sets with unattested Latin proto-words extracted from the test dataset (Reinheimer Ripeanu, 2001).

4.2. Results and Discussion

Since the dataset includes the unattested Latin words (the gold standard), we were able to perform an automatic evaluation of our system's performance. First, we evaluated the individual systems, trained for each modern language independently. Then, we applied and evaluated the ensemble-based aggregation method described in Section 2. The aggregation method obtained 26.9% top-10 accuracy. The best individual result was obtained by the system trained on Italian, which obtained 38% top-10 accuracy, but out of a subset of only 39 unattested words (those for which the Italian cognate was present), so this would mean that for $\tilde{1}5$ Italian words the correct Latin word was found in the first 10 productions; this shows that the forms are confirmed by our method for more than a quarter of the unattested Latin words that were artificially reconstructed. In Table 5 we provide examples of reconstructing unattested Latin words from Romance languages.

5. Conclusions and Future Work

In this paper we proposed a computational approach for two problems in historical linguistics: producing missing cognates with Latin etymology from Romance languages (with focus on Romanian) and reconstructing unattested Latin words.

As future work, we intend to develop this study on several directions.

Phonetic versus graphic. Of the Romance languages, only contemporary French presents differences – often considerable – between the graphic and the phonetic form of the huge majority of words. Contemporary French cognates will have to be approached in their phonetic form. Furthermore, we intend to compare them with their older versions (Old French used a predominantly phonetic spelling). An

important support is the recourse to the words with French and Latin origin from the English vocabulary (representing about 60% of the total English vocabulary).

From phonetics to morphology. The computational approach currently operates with the lemmatized forms of words (the nominative nouns, the masculine adjectives, the infinitive of verbs). We are considering extending this approach to the entirety of the paradigms. In this way, for Romanian we could find out how, for example, the strong or weak conjugations of the verbs were selected, how certain nouns passed from one gender to another, and so on.

Romanian and neighboring languages. The proposed system can obtain forms that, non-existent in contemporary Romanian, may or may not have existed, as ancient borrowings, in neighboring language spaces (such as in Bulgarian, Serbian, Hungarian).

Towards semantics. Finally, we plan to prepare the stage in which the analysis will also tackle the semantic dimension of the vocabulary.

Our results show how the tool we propose can be used for assisting domain experts in studying the evolution and reconstruction of the languages.

6. Bibliographical References

Bhargava, A. and Kondrak, G. (2009). Multiple Word Alignment with Profile Hidden Markov Models. In *Proceedings of NAACL-HLT 2009, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 43–48.

Bouchard-Côté, A., Griffiths, T. L., and Klein, D. (2009). Improved Reconstruction of Protolanguage Word Forms. In *Proceedings of NAACL 2007*, volume 7, pages 65–73.

Campbell, L. (1998). *Historical Linguistics. An Introduction*. MIT Press.

Ciobanu, A. M. and Dinu, L. P. (2014). Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of LREC 2014*, pages 1038–1043.

Ciobanu, A. M. and Dinu, L. P. (2018). Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1604–1614.

Ciobanu, A. M. and Dinu, L. P. (2019). Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Fischer, I. (1985). *Latina Dunăreană. Introducere în istoria limbii române*. Editura Ştiinţifică şi Enciclopedică, Bucureşti.

Fox, A. (1995). *Linguistic Reconstruction*. Oxford University Press, Oxford.

Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000*, pages 288–295.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*, pages 282–289.

List, J.-M., Greenhill, S. J., and Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18, 01.

List, J. (2019). Automatic Inference of Sound Correspondence Patterns across Multiple Languages. *Computational Linguistics*, 45(1):137–161.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453.

Rama, T., List, J., Wahle, J., and Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 393–400.

Reinheimer Ripeanu, S. (2001). *Lingvistica Romanică: Lexic, Morfologie, Fonetică*. Editura All, Bucureşti.

Sala, M. (1998). *De la Latină la Română*. Editura Univers Enciclopedic, Bucureşti.

Weiss, M., (2015). *The Routledge Handbook of Historical Linguistics*, chapter The comparative method. Routledge, New York.